Virology 493 (2016) 142–153

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa

Pauline Bernardo [a,1], Brejnev Muhire [b], Sarah François [a,c,d], Maëlle Deshoux [a], Penelope Hartnady [b], Kata Farkas [e], Simona Kraberger [e], Denis Filloux [a], Emmanuel Fernandez [a], Serge Galzi [a], Romain Ferdinand [a], Martine Granier [a], Armelle Marais [f,g], Pablo Monge Blasco [h], Thierry Candresse [f,g], Fernando Escriu [h,i], Arvind Varsani [e,j,k], Gordon W. Harkins [l], Darren P. Martin [b], Philippe Roumagnac [a,*]

[a] CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France
[b] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa
[c] INRA, UMR 1333, DGIMI, Montpellier, France
[d] CNRS-IRD-UM1-UM2, UMR 5290, MIVEGEC, Avenue Agropolis, Montpellier, France
[e] School of Biological Sciences and Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
[f] INRA, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France
[g] Université de Bordeaux, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France
[h] Unidad de Sanidad Vegetal, Centro de Investigacion y Tecnologia Agroalimentaria de Aragon (CITA), Av. Montañana 930, 50059 Zaragoza, Spain
[i] Unidad de Sanidad Vegetal, Instituto Agroalimentario de Aragón IA2 (CITA - Universidad de Zaragoza), Av. Montañana 930, 50059 Zaragoza, Spain
[j] Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, USA
[k] Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Observatory, South Africa
[l] South African National Bioinformatics Institute, MRC Unit for Bioinformatics Capacity Development, University of the Western Cape, Cape Town, South Africa

ARTICLE INFO

Article history:
Received 20 December 2015
Returned to author for revisions
22 March 2016
Accepted 23 March 2016

Keywords:
Geminiviridae
Alfalfa leaf curl virus
Euphorbia caput-medusae latent virus
Prevalence
Recombination
Secondary structure
Genome organization
France
Spain
South Africa

ABSTRACT

Little is known about the prevalence, diversity, evolutionary processes, genomic structures and population dynamics of viruses in the divergent geminivirus lineage known as the capulaviruses. We determined and analyzed full genome sequences of 13 Euphorbia caput-medusae latent virus (EcmLV) and 26 Alfalfa leaf curl virus (ALCV) isolates, and partial genome sequences of 23 EcmLV and 37 ALCV isolates. While EcmLV was asymptomatic in uncultivated southern African *Euphorbia caput-medusae*, severe alfalfa disease symptoms were associated with ALCV in southern France. The prevalence of both viruses exceeded 10% in their respective hosts. Besides using patterns of detectable negative selection to identify ORFs that are probably functionally expressed, we show that ALCV and EcmLV both display evidence of inter-species recombination and biologically functional genomic secondary structures. Finally, we show that whereas the EcmLV populations likely experience restricted geographical dispersion, ALCV is probably freely moving across the French Mediterranean region.

1. Introduction

Next Generation Sequencing and metagenomics-based study designs have impacted our appreciation of the prevalence, pervasiveness and diversity of environmental single stranded DNA (ssDNA) viruses (Candresse et al., 2014; Filloux et al., 2015b; Kraberger et al., 2015; Ng et al., 2014, 2009, 2011; Roossinck et al., 2015; Rosario and Breitbart, 2011). Among the best studied of these ssDNA viruses have been the plant-infecting viruses in the family *Geminiviridae*. The past four decades have seen the worldwide emergence of several major plant diseases caused by geminiviruses and also the discovery of hundreds of previously unknown, and sometimes highly divergent, geminiviral species (Bernardo et al., 2013; Briddon et al., 2010; Liang et al., 2015; Loconsole et al.,

* Corresponding author.
E-mail address: philippe.roumagnac@cirad.fr (P. Roumagnac).
[1] P.B. and B.M. contributed equally to this work.

http://dx.doi.org/10.1016/j.virol.2016.03.016
0042-6822/

2012; Ma et al., 2015; Roumagnac et al., 2015; Varsani et al., 2009; Yazdi et al., 2008). The rate at which new geminiviruses are being discovered has recently been accelerated through the development and application of sequence-non-specific virus discovery approaches (Roossinck et al., 2015; Rosario et al., 2012) such as rolling circle amplification (RCA) based virus cloning and sequencing (Haible et al., 2006; Inoue-Nagata et al., 2004; Shepherd et al., 2008) and VANA (virion-associated nucleic acids) or siRNA based metagenomics (Candresse et al., 2014; Filloux et al., 2015a; Kreuze et al., 2009). In studies exploring the diversity of these viruses, such approaches have facilitated the expansion of sampling efforts to include both insects (both virus vectors and their predators) and uncultivated host species (Bernardo et al., 2013; Ng et al., 2011; Rosario et al., 2011, 2015). The application of these improved sampling and virus discovery strategies have revealed that geminivirus diversity exceeds that which is currently known (Haible et al., 2006; Ng et al., 2011; Schubert et al., 2007). Such studies have also led to a reevaluation of the known geographical ranges of the different geminivirus genera with, for example, the overturning of the long-held view that members of the genus *Mastrevirus* do not occur in the Americas (Agindotan et al., 2015; Candresse et al., 2014; Kreuze et al., 2009).

Since some of the geminiviruses discovered over the past few years are highly divergent, and, in some cases, have unique genome architectures (Briddon et al., 2010; Loconsole et al., 2012; Varsani et al., 2009; Yazdi et al., 2008), three new geminivirus genera were created and approved in 2014 (*Becurtovirus*, *Turncurtovirus*, and *Eragrovirus*; Varsani et al., 2014b) and at least three others are likely to follow (Bernardo et al., 2013; Krenz et al., 2012; Loconsole et al., 2012); among which are the capulaviruses (Bernardo et al., 2013). Three distinct species of this new lineage were discovered between 2010 and 2011 infecting, respectively, a wild spurge, *Euphorbia caput-medusae* in South Africa (Euphorbia caput-medusae latent virus, EcmLV; Bernardo et al., 2013), alfalfa (*Medicago sativa*) in France (Alfalfa leaf curl virus, ALCV; Roumagnac et al., 2015), and French bean (*Phaseolus vulgaris*) in India (French bean severe leaf curl virus, FbSLCV; Accession number NC_018453). In addition to their high degree of sequence divergence, these viruses also exhibit a genome organization that is unique among the geminiviruses (Bernardo et al., 2013; Roumagnac et al., 2015). ALCV and FbSLCV cause severe symptoms in alfalfa and French bean, respectively, and it has recently been shown that ALCV is transmitted by *Aphis craccivora* (Roumagnac et al., 2015); an invasive aphid species with an almost global distribution (CIE, 1983).

We here collected hundreds of alfalfa and *E. caput-medusae* plants from France and South Africa, respectively, and comparatively analyze the genome sequences of 13 isolates of EcmLV and 26 ALCV along with that of FbSLCV. We show that both EcmLV and ALCV have high prevalence (12–13%) in their respective host species in the Western Cape region of South Africa and in three Southern regions of France. We also present a new codon-model based natural selection detection approach to reveal open reading frames that are probably functionally expressed in capulaviruses genomes. We further demonstrate that, as with other geminiviruses, capulavirus genomes display evidence of both inter-species recombination and biologically functional secondary structures.

2. Materials and methods

2.1. Plant sampling

In 2014, 238 *M. sativa* plants were randomly collected (i.e. irrespective of the presence of potential symptoms) from three regions of Southern France, including the Rhône delta (seven sampling locations), the Montpellier region (four sampling locations) and the Toulouse region (two sampling locations; Supplementary Fig. 1). The symptom status of the 238 plants was assessed and plants were then stored at −80 °C prior to virus detection and characterization. A further 43 symptomatic and 15 asymptomatic *M. sativa* plants were collected later in 2014 from the same three areas (Supplementary Fig. 1). An additional four symptomatic alfalfa plants collected in 2012 and 2013 from the Ebro valley region of the Zaragoza province of Spain were also included for further analysis. Collectively, 300 alfalfa plants were obtained from France and Spain between 2012 and 2014.

In 2012, 302 asymptomatic *Euphorbia caput-medusae* were randomly collected from seven separate locations within the Western Cape region of South Africa (Supplementary Fig. 1). These samples were stored at −80 °C. In 2015, 14 additional samples were collected from an eighth location at the University of the Western Cape Nature Reserve (Supplementary Fig. 1).

2.2. DNA extraction, amplification, cloning and sequencing

Total DNA from alfalfa and *E. caput-medusae* plant samples was extracted as previously described by Bernardo et al. (2013).

PCR-mediated detection of ALCV from the 296 alfalfa plants collected in France, and four plants from Spain was performed using two pairs of PCR primers (ALCV-187F forward primer 5'-TGG AAT ATT GTG CTG CTT GG-3' and ALCV-971R reverse primer 5'-ATT TTG GGA CTT GTG CTC CA-3'; and ALCV-986F forward primer 5'-ATG ATG GAT AAT TCA AAC CC-3' and ALCV-1202R reverse primer 5'-TTC TTC TGG GTA TTT GCA TA-3'). Amplification conditions consisted of 94 °C for 2 min; 30 cycles at 94 °C for 1 min, 58 °C for 1 min (primer pair 1)/55 °C for 30 s (primer pair 2), 72 °C for 50 s; and 72 °C for 5 min. Amplicons were directly sequenced using automated Sanger sequencing (Beckman Coulter Genomics). Circular DNA molecules from samples that tested positive by at least one of the two PCR assays were enriched using RCA (using TempliPhi™, GE Healthcare, USA) as previously described (Shepherd et al., 2008). The RCA products were used as a template for PCR using an abutting pair of primers designed from the 44-1E ALCV complete genome (Accession number KP732474; Roumagnac et al., 2015); Cap-ncoIF: 5'-CCA TGG CCT TCA AAG GTA GCC CAA TTC AAY ATG G-3' and Cap-ncoIR: 5'-CCA TGG GGC CTT ATY CCT CKG YGA TCG-3' using KAPA HiFi Hotstart DNA polymerase (Kapa Biosystems, USA). Amplification conditions consisted of: 96 °C for 3 min, 25 cycles at 98 °C for 20 s, 60 °C for 30 s, 72 °C for 165 s, and 72 °C for 3 min. The amplicons were gel purified, cloned into pJET2.1 (Thermo Fisher, USA) and Sanger sequenced by primer walking at Macrogen Inc. (Korea). In addition, RCA products were digested with *Eco*RI, *Bam*HI, *Dra*I, *Nco*I or *Nde*I for 3 h at 37 °C in order to screen for the presence of a potential DNA-B geminiviral component or satellite sequences.

PCR-mediated detection of EcmLV from the 316 *E. caput-medusae* plants was performed using two pairs of PCR primers: (i) Dar-136F forward primer 5'-CGA AGA GGT CAT TGG GAC AT-3' and Dar-730R reverse primer 5'-CGG GTC TGG CTA AGA GAG TG-3' as previously described by Bernardo et al. (2013) and (ii) Dar-1775F forward primer 5'-TTG AAT TGC ATG GGC ACT TA-3' and Dar-2433R reverse primer 5'-GCC CTT TTG GTC ATT TTG AA-3'. Amplification conditions consisted of: 95 °C for 5 min; 30 cycles at 94 °C for 1 min, 56 °C for 1 min, 72 °C for 50 s; and 72 °C for 5 min. Circular DNA molecules from samples that tested positive by at least one of the two PCR assays was enriched using RCA as described above for alfalfa. RCA products were all digested with *Eco*RI for 3 h at 37 °C. Subsequently, samples that could not be cleaved using *Eco*RI were digested with *Bam*HI for 3 h at 37 °C. Geminivirus-like genomes from 13 *E. caput-medusae* samples were cloned in pGEM-T Easy (Promega Biotech) using methods described by Bernardo et al. (2013).

Prevalence was defined as the proportion of alfalfa or *E. caput-medusae* plants being infected by ALCV or EcmLV from the alfalfa or *E. caput-medusae* populations that were randomly collected in France or South Africa, respectively.

2.3. Sequence analysis

Sequence contigs were assembled using BioNumerics Applied Maths V6.5 (Applied. Maths, Ghent, Belgium) and were compared to sequences in the GenBank database using BLASTn and BLASTx (Altschul et al., 1990). All pairwise identity analyses of the full genome nucleotide sequences, capsid protein (CP) amino acid sequences, and replication associated protein (Rep) amino acid sequences were carried out using the MUSCLE-based pairwise alignment (Edgar, 2004) option implemented in SDT v1.2 (Muhire et al., 2014b).

2.4. Detection of conserved secondary-structural elements within capulavirus genomes

The computer program Nucleic Acid Secondary Structure Predictor (NASP) (Semegni et al., 2011) was used as previously described (Muhire et al., 2014a) to identify the conserved secondary-structural elements present within 45 capulavirus genomes (EcmLV, $n=16$; ALCV, $n=27$; FbSLCV, $n=2$). In each of the data sets, secondary-structural elements were first inferred using a minimum free-energy (MFE) approach implemented in hybrid-ssmin (a component of the UNAFold package; (Markham and Zuker, 2008). From amongst sets of plausible whole genome secondary structures (approximately ~10 alternative folds per genome) NASP identified subsets of conserved high-confidence structural elements - referred to as high-confidence structure sets (HCSSs). For the NASP analysis, sequences were folded as circular ssDNA at 25 °C under 1 M sodium. In subsequent analyses, only nucleotides identified as being base-paired within the HCSSs were treated as paired sites, whereas all other nucleotides were treated as unpaired sites. The ALCV and EcmLV datasets contained enough sequences that were sufficiently divergent to test for evidence of evolutionary pressures favoring the maintenance of base–pairing interactions within the HCSSs (Muhire et al., 2014a). Three different tests, all designed to test whether sequences were evolving in a way consistent with the evolutionary preservation of biologically functional structural elements, were applied exactly as recently described (Muhire et al., 2014a). First, at the whole-genome-scale an allele frequency spectrum permutation test (Fu and Li, 1993; Tajima, 1989), which compares frequencies of alternative alleles at paired vs unpaired sites, was used to compare degrees of negative selection (selection disfavoring change) at paired vs unpaired sites. Second, within the CP and Rep gene coding regions, a Maximum Likelihood codon-model based test (based on the FUBAR method (Murrell et al., 2013) implemented in HyPhy (Pond et al., 2005) was used to compare synonymous substitution rates in codons containing paired and unpaired third codon position nucleotides. Third, the complementary coevolution detection method (Muhire et al., 2014a) (based on the SPIDERMONKEY method (Poon et al., 2008), also implemented in HyPhy) was used to test whether pairs of nucleotides that were base-paired within the HCSSs also displayed any evidence of complementary coevolution (i.e. coevolution specifically favoring the maintenance of base-pairing). Structures were visualized and, based on the various analyses performed, ranked in order of their likely biological functionality using the computer program, DOOSS (http://dooss.computingforbiology.org; Golden and Martin, 2013).

2.5. Identification of open reading frames (ORFs) and capulavirus genome organizations

Identification of open reading frames (ORFs) was initially performed using the ORF Finder ncbi graphical analysis tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). In addition, we devised a computational tool, named "ORFunc" that detects all ORFs above a user defined length that are both conserved within a multiple sequence alignment (in this case an alignment of all available capulavirus sequences) and are evolving under a detectable degree of purifying selection (i.e. selection that disfavoring non-synonymous substitutions). ORFunc ranks ORF in order of their likely functional expression based on the degree of detected purifying selection against changes in the potentially encoded amino acid sequence. Specifically, given a multiple sequence alignment in FASTA format, ORFunc performs a preliminary scan of the alignment to generate a list of unique ORFs sharing greater than a predetermined pairwise sequence identity (in the present analyses 0.80 to ensure the sequences within each of the resulting sub-alignments were credibly aligned) and above a given sequence length (in the present analyses 100nt; a size slightly below that of any currently known geminivirus gene). A local BLAST search (Altschul et al., 1990) is performed using the command-line version of NCBI-BLAST (available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/) whereby each ORF is used as a query to search a local BLAST database produced from the input alignment so as to obtain matches for other sequences: matches that are then used to produce a codon alignment. For each codon alignment thus produced, synonymous substitution rates (dS), non-synonymous substitution rates (dN) and the probability of negative selection ($dS > dN$) at each codon site are estimated using the FUBAR method (Murrell et al., 2013). This approach enabled the determination of whether inferred ORFs within the various EcmLV and ALCV genomes were evolving under negative selection (identified by their constituent codons on average displaying significantly higher synonymous substitution rates than non-synonymous substitution rates). ORF-wide probabilities of negative selection are then estimated by combining the p-values of non-negative selection at each individual codon-site (Theiler and Bloch, 1996). To test the reliability of this approach, two geminivirus datasets for which genes have already been well-characterized (*Tomato yellow leaf curl virus* [TYLCV] and *Maize streak virus* [MSV]; Dry et al., 1993; Lazarowitz et al., 1989) were used as controls. ORFunc is written in python and is available from: web.cbio.uct.ac.za/~brejnev/downloads/ORFunc.tar.gz

2.6. Analysis of potential recombination events

Evidence of potential recombination events was detected within the 45 capulavirus full-genome alignment using the RDP, GENECONV, BOOTSCAN, MAXIMUM CHI SQUARE, CHIMAERA, SISCAN and 3SEQ recombination detection methods that are implemented in RDP4.50 with default settings (Martin et al., 2015). Only recombination events detected with three or more distinct groups of methods (where RDP and GENECONV were each considered as distinct methods and BOOTSCAN/SISCAN and MAXIMUM CHI SQUARE/CHIMAERA/3SEQ were considered as single distinct method groups) and that also had significant phylogenetic support, were considered credible evidence of recombination.

2.7. Phylogenetic analysis

The evolutionary relationships of capulaviruses and other geminiviruses were reconstructed using Rep and CP amino acid sequences (i.e. the only proteins that are clearly homologous across all geminiviruses). Datasets consisting of 45 predicted capulavirus Rep and CP amino acid sequences together with the corresponding homologous sequences from Grapevine cabernet franc-associated virus (GCFaV, JQ901105; Krenz et al., 2012) chosen as a divergent non-capulavirus outlier used to root the Rep and CP phylogenies. Predicted Rep and CP amino acid sequences were aligned using MUSCLE. Maximum likelihood phylogenetic trees of the Rep and CP were inferred using PhyML3 (Guindon et al., 2009) implemented in MEGA with the rtREV+G+F amino acid substitution model chosen as the best-fit using ProtTest (Abascal et al., 2005). Five hundred bootstrap replicates were used to test the support of branches. In addition, 64 ALCV gene fragments (ALCV 44-1 genomic positions 260–806) 547 nt in length encompassing the V3 ORF and part of the

cp ORF were aligned using the MUSCLE method (Edgar, 2004) implemented in MEGA (with default settings). A maximum likelihood phylogenetic tree was constructed using PhyML3 with a K2+G+I nucleotide substitution model (selected as best fit by MEGA) and 500 bootstrap replicates were used to test the support of branches. Finally, all 45 currently available whole capulaviruses genome sequences were aligned using MUSCLE. A maximum likelihood phylogenetic tree was constructed using PhyML3 with a TN93+G nucleotide substitution model (selected as best fit by MEGA) with 500 bootstrap replicates used to test the support of branches.

2.8. Statistical analyses

Mantel (1967) tests conducted using XLSTAT (10.000 permutations) were used to test for evidence of correlation between genetic and geographic distances for 39 × 508 nt long EcmLV gene fragments (EcmLV-Dar10 genomic positions 1877–2384) encompassing the C3 ORF and part of the C1 ORF and for 41 × 547 nt long ALCV gene fragments (ALCV 44-1 genomic positions 260–806) encompassing the V3 ORF and part of the *cp* ORF. Samples collected at the regional scale (Southern regions of France for ALCV and Western Cape region for EcmLV) were used (Supplementary Fig. 1). The genetic distance matrix was obtained using MEGA 5.2.1 (CLUSTALW alignment followed by uncorrected pairwise distance estimation using the pairwise deletion option) and the geographic distance matrix was obtained using the program Geographic Distance Matrix Simulator 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg).

3. Results and discussion

3.1. Characterization of a collection of ALCV isolates from France

Based on the discovery of ALCV in 2010 (Roumagnac et al., 2015), broad sampling surveys in three regions of Southern France and in one region of Spain were conducted during 2012–2014 (Supplementary Fig. 1). PCR analysis of the collected samples revealed that ALCV was present in all four regions. Because the virus was detected in 32/238 plants that were randomly collected in France the percentage of infected plants (13.4%) is likely a valid estimate of the prevalence of ALCV in alfalfa across the three French regions (Supplementary Fig. 1). The prevalence of ALCV at each of the thirteen French sampling sites ranged from 1.2% (La Tour du Valat, Rhône delta region) to 45.8% (Petit Bastières, Rhône delta region).

Visual comparisons of the 32 ALCV-positive plants with plants that tested negative revealed that, relative to non-infected plants, all ALCV-infected plants were stunted and consistently displayed varying degrees of leaf curling, crumpling and shriveling (Supplementary Fig. 2). These potential symptoms unequivocally resemble those observed in alfalfa plants infected by aphid-inoculation with the 44–1E infectious clone of ALCV (Roumagnac et al., 2015). It is noteworthy that enations such as those observed in 44–1E agroinoculated faba beans (Roumagnac et al., 2015) were neither observed in the sampled plants, nor in 44–1E aphid-inoculated alfalfa plants (Roumagnac et al., 2015). Based on the presence or absence of these conspicuous symptoms, an additional 43 symptomatic and 15 asymptomatic plants were collected later in 2014 from the three previously sampled regions of Southern France (Supplementary Fig. 1). A diagnostic PCR detected ALCV in 39/43 of the symptomatic plants but in none of the asymptomatic plants, suggesting that the symptoms observed in the field are likely caused by ALCV infection.

Twenty-six ALCV complete genome sequences were obtained that ranged in size from 2737 nt to 2769 nt in length and shared > 82.5% genome-wide pairwise identity (Supplementary Fig. 3). This degree of similarity is above the species demarcation thresholds recommended for all of the geminivirus genera (Muhire et al., 2013; Varsani et al., 2014a, 2014b) except for the begomoviruses (which have a species demarcation threshold of 91% (Brown et al., 2015) showing that the 26 isolates from which these genomic sequences were obtained could be reasonably, albeit tentatively, classified as ALCV variants. The circular DNA molecules obtained using RCA were tentatively considered to be the complete genomes of geminiviruses infecting the alfalfa plants, because only one band was resolved by electrophoresis of the digested RCA products.

3.2. Characterization of a collection of EcmLV isolates from South Africa

In order to both assess the diversity of EcmLV genome sequences, and determine the prevalence of EcmLV within the Western Cape province of South Africa, 316 asymptomatic *E. caput medusae* plants from eight dispersed locations were sampled in this region (Supplementary Fig. 1). Using a diagnostic PCR, EcmLV was detected in 38/316 plants sampled in 5/8 of the sampling locations. This indicates an overall EcmLV prevalence of 12% at the Western Cape regional scale: a prevalence very similar to that of 13.4% found for ACLV in alfalfa in Southern France. The local prevalence of EcmLV ranged from 5% (at the Silver Stream sampling site) to 30% (at the Paternoster site).

The 13 EcmLV complete genome sequences that were obtained together with three full genomes previously described (Bernardo et al., 2013) had pairwise identities ranging between 92.8 and 99.7% (Supplementary Fig. 3), confirming that all isolates belong to the same species (Bernardo et al., 2013). As for ALCV, the circular DNA molecules obtained using RCA were tentatively considered to be the complete genomes of geminiviruses infecting the *E. caput-medusae* plants, because only one band was resolved by electrophoresis of the digested RCA products.

3.3. Characterization of capulavirus genome organizations

ORFunc was used to identify the potential functional genes of EcmlV and ALCV, with the well-known geminiviruses *Maize streak virus* (MSV) and *Tomato yellow leaf curl virus* (TYLCV) being used as controls for the validation of this new tool. ORFunc is designed to detect and rank ORFs in order of the likelihood of their functional expression based on overall evidence of negative selection detected across all of their constituent codon sites. It is expected that the majority of functional ORFs should be evolving under some degree of negative selection favoring the maintenance of functionally important amino acid sequences. By implementing such a selection-based approach, ORFunc goes beyond simply annotating genomic fragments delimited by start and stop codons. As a proof-of-concept, ORFunc consistently identified significant degrees of negative selection across all four known functional genes of MSV (Fig. 1 and Supplementary Dataset 1) and the *rep* (C1), transcription activator protein (C2), replication enhancer protein (C3) and coat protein (V1) genes of TYLCV (Fig. 1 and Supplementary Dataset 1). ORFunc did not identify any potentially functional genes other than those illustrated in Fig. 1. In addition, the failure to obtain statistically significant evidence of negative selection in the V2 and C4 ORFs in TYLCV should not be equated with the absence of negative selection in these ORFs. The C4 gene is completely contained within the *rep* gene, which means that very few possible substitutions in this ORF would be synonymous (i.e. almost all possible substitutions that would not change the encoded amino acid sequence encoded by C4 would change the amino acid sequence of Rep and would, therefore, not be synonymous). In the case of V2 whereas 10/134 individual codon sites are apparently evolving under a significant degree of negative selection (with a p-value cutoff of 0.1), 1/134 codon sites are apparently evolving under a significant degree

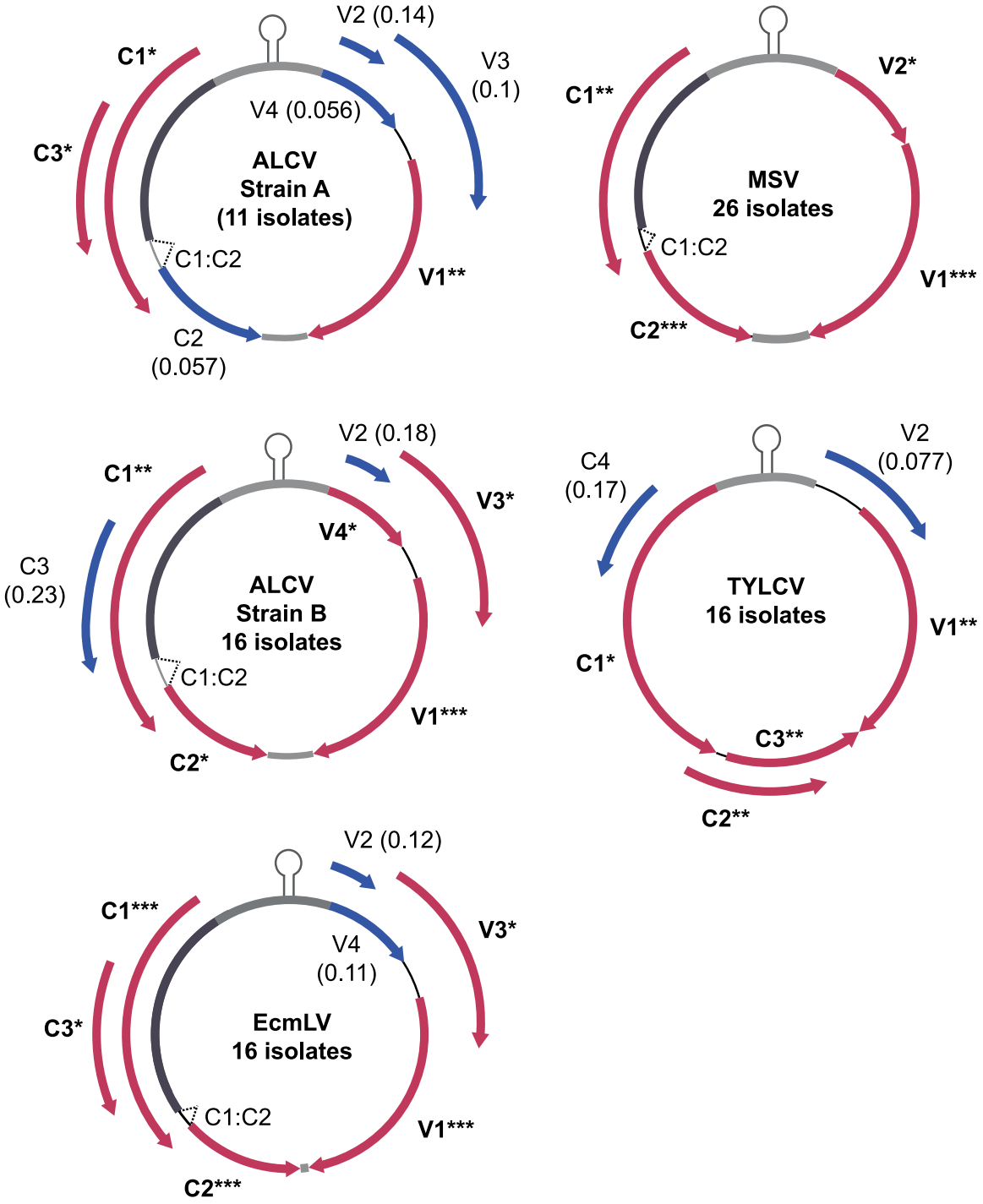

Fig. 1. Genomic organization of ALCV strain A, ALCV strain B, EcmLV, MSV and TYLCV showing the arrangements of potential genes. Open reading frames (ORFs) larger than 100 nucleotides that contain predicted codon sites that are collectively evolving under significant degrees of negative selection (i.e. for which ORF-wide estimates of synonymous substitution rates are significantly higher than non-synonymous substitution rate estimates) are indicated in pink ($^{*}=p$-value <0.05, $^{**}=p$-value <0.01, and $^{***}=p$-value <0.001; see Supplementary Dataset 1). ORFs containing predicted codons that are not evolving under significant degrees of negative selection ($p>0.05$) are indicated in blue (with the associated p-value for the synonymous substitution rate being lower than or equal to the non-synonymous rate indicated in brackets; see Supplementary Dataset 1).

of positive selection. Most importantly, however, ORFunc yielded no false-positive evidence of functionally expressed ORFs in MSV and TYLCV (Fig. 1 and Supplementary Dataset 1).

Of the seven large ORFs identified by the ncbi ORF Finder graphical analysis tool that are apparently conserved between the various capulavirus genomes (V1–V4 and C1–C3; Fig. 1), three (V1, C1 and C3), five (V1, V3, V4, C1 and C2) and five (V1, V3, C1, C2 and C3) were detectably evolving under significant ORF-wide negative selection (associated p-value <0.05) in ALCV strain A (see below regarding ALCV strain A and B demarcation), ALCV strain B, and EcmLV, respectively (Fig. 1 and Supplementary Dataset 1). The V2 ORF did not display significant ORF-wide evidence of negative selection in any of the three groups, suggesting that this ORF might not be functionally expressed. It should be noted, however, that absence of a significant negative selection signal is not proof that this ORF is not functionally expressed. Bigger and more

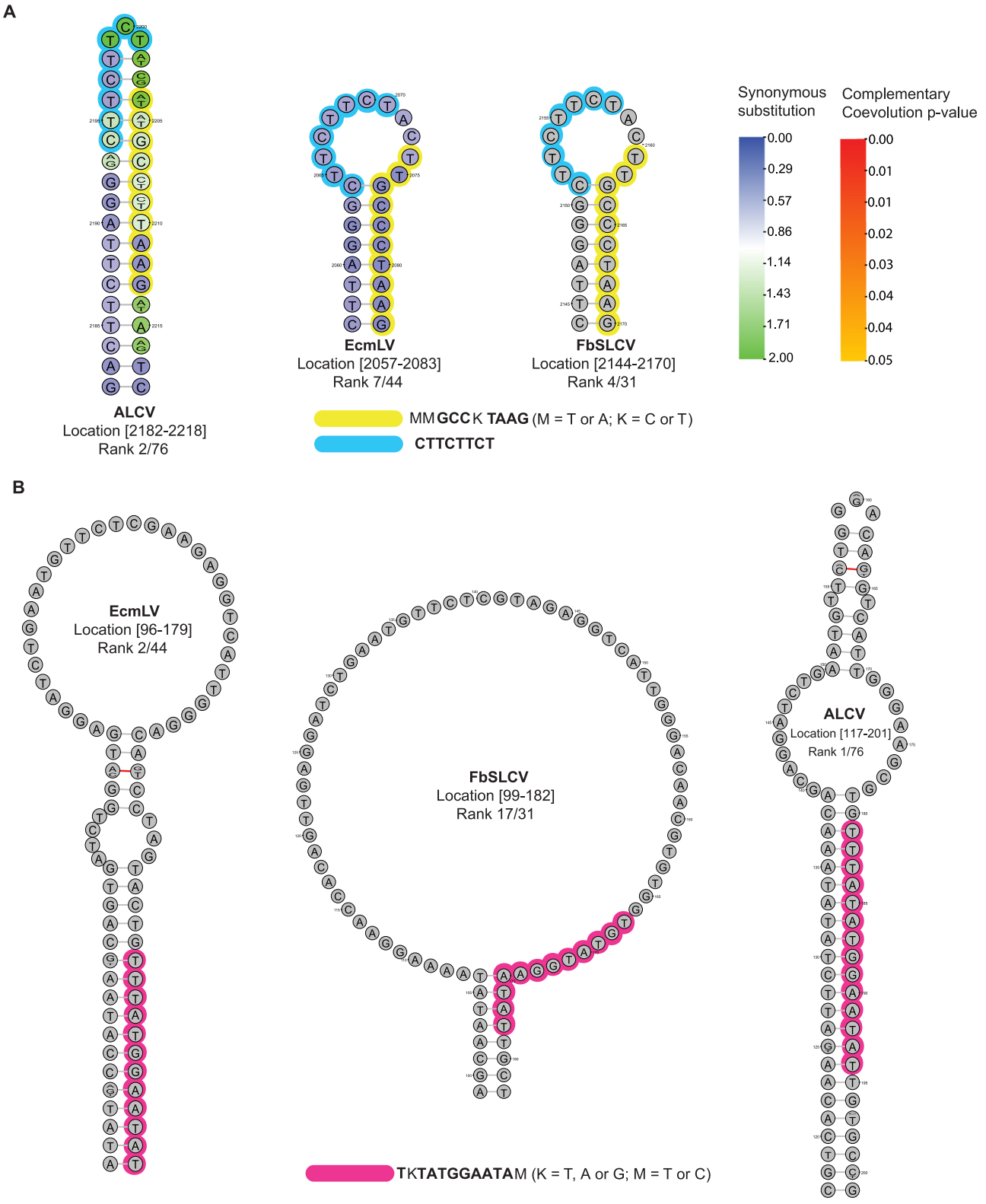

Fig. 2. (A) Secondary structures associated with the 5' end of capulavirus C2 ORFs. The similarities between these structures include homologous stem-loop sequences (highlighted in yellow), and a highly conserved sequence motif found in the ALCV, EcmLV and FbSLCV data sets (highlighted in blue). This structure is ranked highly within the high confidence structure sets of all data sets (seventh out of 44 for EcmLV, second out of 76 for ALCV and fourth out of 31 for FbSLCV). Nucleotide sequence variability at each nucleotide site is reflected by a sequence logo at each position. For the ALCV and EcmLV structures substitution rate estimates are represented by the color of nucleotide triplets falling within individual predicted codon sites (ranging from blue to green). Low synonymous substitution rates are observable in the stem region, which is consistent with a high degree of evolutionary conservation of the nucleotides comprising these codons. Nucleotides falling outside of the gene or for which synonymous substitution rates could not be estimated due to too few sequences being available for analysis (as in the FbSLCV), are shaded gray. (B) Capulavirus secondary structure predicted in the long intergenic region (LIR). The AT-rich stem-loop structure, with a 12 nt conserved sequence (highlighted in pink), is highly ranked in all of the analyzed capulavirus datasets (first out of 76 HCSS structures for ALCV, second out of 44 in EcmLV and 17th out of 31 for FbSLCV). In the case of ALCV and EcmLV, base-paired sites displaying significant degrees of complementary coevolution (P value <0.05) are indicated by the red lines between the nucleotides in the stem region of each structure. It is plausible that this highly conserved structural element either contains the complementary strand origin of replication or is involved in the regulation of virion and/or complementary gene expression.

diverse capulavirus genomic sequence datasets will substantially increase the power of ORFunc to detect low degrees of negative selection within this ORF: possibly to the point where it is able to confirm the functional expression of V2.

3.4. Detection of conserved secondary-structural elements within capulavirus genomes

Recent computational analyses of geminiviruses in the genera *Mastrevirus* and *Begomovirus* have revealed evidence of evolutionarily conserved (and hence likely biologically functional) genomic secondary structures (Muhire et al., 2014a). Using NASP, we identified 76, 44, and 31 high-confidence structural elements within the ALCV, EcmLV, and FbSLCV genome datasets, respectively.

Besides a conserved stem-loop structure at the presumed virion-strand origin of replication that resembles those found in other geminiviruses (Lazarowitz, 1992), a further 12 uncharacterized but potentially biologically functional genomic and/or mRNA structural elements that are clearly conserved across all three of the analyzed capulavirus species were identified. Amongst these are two particularly interesting uncharacterized structural elements that display high degrees of evolutionary conservation across both the capulaviruses and geminiviruses in other genera (Fig. 2). The first of these structural elements, which is highly ranked within the HCSSs of all three capulavirus species, is a conserved hairpin-loop structure with a 10–20 nt long stem within the C2 ORF (Fig. 2). This structure is possibly homologous to a similarly situated secondary structure previously identified using these same computational methods at the 5' end of MSV C2 ORF.

The second particularly conserved new structural element identified is another hairpin located within the long intergenic region (LIR) of all the analyzed capulavirus species. It contains an A-T rich sequence (Fig. 2) and resembles a structure previously identified in two different mastrevirus species (MSV and *Panicum streak virus*). In diverse ssDNA viruses, virion strand origins of replication (*v-oris*) consist of hairpin structures with highly conserved AT-rich loop sequences that generally occur within intergenic regions (IRs). It is plausible that this second conserved structural element may be either associated with the as yet undiscovered capulavirus complementary strand origin of replication (which in begomoviruses is close to the *v-ori*), or to the regulation of transcription and replication (both of which are known to be regulated by sequence elements within the IRs of various other geminiviruses; Arguello-Astorga et al., 1994; Gutierrez et al., 2004).

It is noteworthy that these analyses also identified what appears to be a repeated 8 bp sequence (with the consensus 5'-AGGCCCAA-3') within the stem regions of multiple structural elements within the HCSSs identified in the various capulavirus datasets. The consensus of the repeated sequence is almost identical to functional sequence motifs previously detected in three other settings. Specifically, it bears a striking resemblance to (i) the 3' ends (3'-AGGCCCA-5') of predicted viral miRNA hairpins (Li et al., 2008); (ii) a regulatory promoter heptamer element involved in the development of plant tissues (5'-AGGCCCA-3') (Obayashi et al., 2007), and (iii) a 7 bp sequence motif (5'-AGGCCCA-3') located upstream of ribosomal protein transcription initiation sites in *Arabidopsis thaliana* (Thompson et al., 1992). It is plausible therefore, that many of the secondary structures identified in these capulavirus genomes (Supplementary Dataset 2) may play a role in modulating the sensitivity of capulavirus genomes to RNA interference (Schubert et al., 2005).

Although their high degree of interspecific evolutionary conservation suggests that the highest ranked of the structural elements identified within the various capulavirus HCCSs are indeed biologically functional, we also tested whether evidence of this biological functionality was apparent within the patterns of nucleotide substitution that the EcmLV and ALCV genomes have undergone (the two capulavirus species with sufficient available data to perform these analyses). Towards this end, the ALCV and EcmLV genomes were partitioned into “paired” and “unpaired” site sets and, focusing only on variable sites (invariant sites were removed), the frequency spectra of the “minor alleles” (i.e. those present at the lowest frequencies within the sampled viruses) were compared at these sites within the paired and unpaired genome partitions. This revealed that while minor allele frequencies were lower at paired sites in both the ALCV and EcmLV genomes (a finding consistent with stronger negative selection acting on paired sites than on unpaired sites), the difference was statistically significant only for EcmLV (permutation test p-value$=0.01$, Fu and Li's F test; Supplementary Table 1). Also consistent with the hypothesis that paired sites are evolving under stronger negative selection than unpaired sites was the detection within both the ALCV and EcmLV Rep and CP coding regions of significantly reduced synonymous substitution rates in codons, which have base-paired third-position nucleotides (respective multiple comparison-corrected Mann–Whitney U test p-values$=0.013$ and 0.032 for ALCV and 0.0005 and 0.0235 for EcmLV; Supplementary Table 2). Furthermore, within the EcmLV dataset a very strong association between nucleotide sites that are complementary coevolving and nucleotide sites that are base-paired (Chi squared p-value$=0.0001096$) was detected (Supplementary Table 3).

Collectively, these results provide evidence that many of the detected structural elements within the EcmLV and ALCV genomes are likely being actively preserved by natural selection, and, in the case of EcmLV at least, that there is substantial evolutionary pressure for the maintenance of specific biologically important base–pairing interactions. Further experimental assays should of course be carried out both to test whether these predicted functional secondary structural elements are indeed functional, and, if they are, to determine the precise aspects of capulavirus biology that they impact.

3.5. Phylogenetic and recombination analyses

Both CP and Rep phylogenetic trees indicate that the new ALCV and EcmLV isolates cluster with the other described capulaviruses isolates (Fig. 3).

Whereas the EcmLV Rep sequences all cluster within a single group of variants sharing $>92.8\%$ genome-wide nucleotide sequence identity (Fig. 3), ALCV Rep sequences cluster within two distinct groups (tentatively referred to here as strain A and strain B), with the isolates in each group differing at $\sim 17.5\%$ of sites relative to those in the other group (Supplementary Fig. 3). Strain A isolates are slightly over-represented among the 64 ALCV isolates for which a 547 nt long fragment encompassing the V3 ORF and part of the *cp* ORF was sequenced (37/64; 58%). Whereas ALCV strain A isolates were found at 11/14 of the sampling sites (including the Spanish one; Supplementary Fig. 4), strain B isolates were only found at 8/14 of the sampling sites (including the Spanish one; Supplementary Fig. 4).

Efforts were made to detect and characterize recombination events that could have occurred among the current dataset of 45 capulavirus full genomes. Fourteen apparently unique recombination events were detected, including one event in the EcmLV genomes and 13 in the ALCV genomes (Table 1, Supplementary Table 4 and Fig. 4). Interestingly, all of the examined ALCV and EcmLV isolates display traces of recombination events, with ALCV isolates displaying, on average, evidence of 2.7 events (Supplementary Table 4).

Three out of the 13 ALCV recombination events apparently involved exchanges of sequences between ALCV variants (events 2, 7 and 14) whereas the other detected events (10 in ALCV and one in EcmLV), apparently involved inter-species sequence transfers (events 1, 3 to 6 and 8–13; Table 1 and Fig. 4). Interestingly, these inter-species sequence transfers all appear to have involved at least one parent that is related to a currently known capulavirus

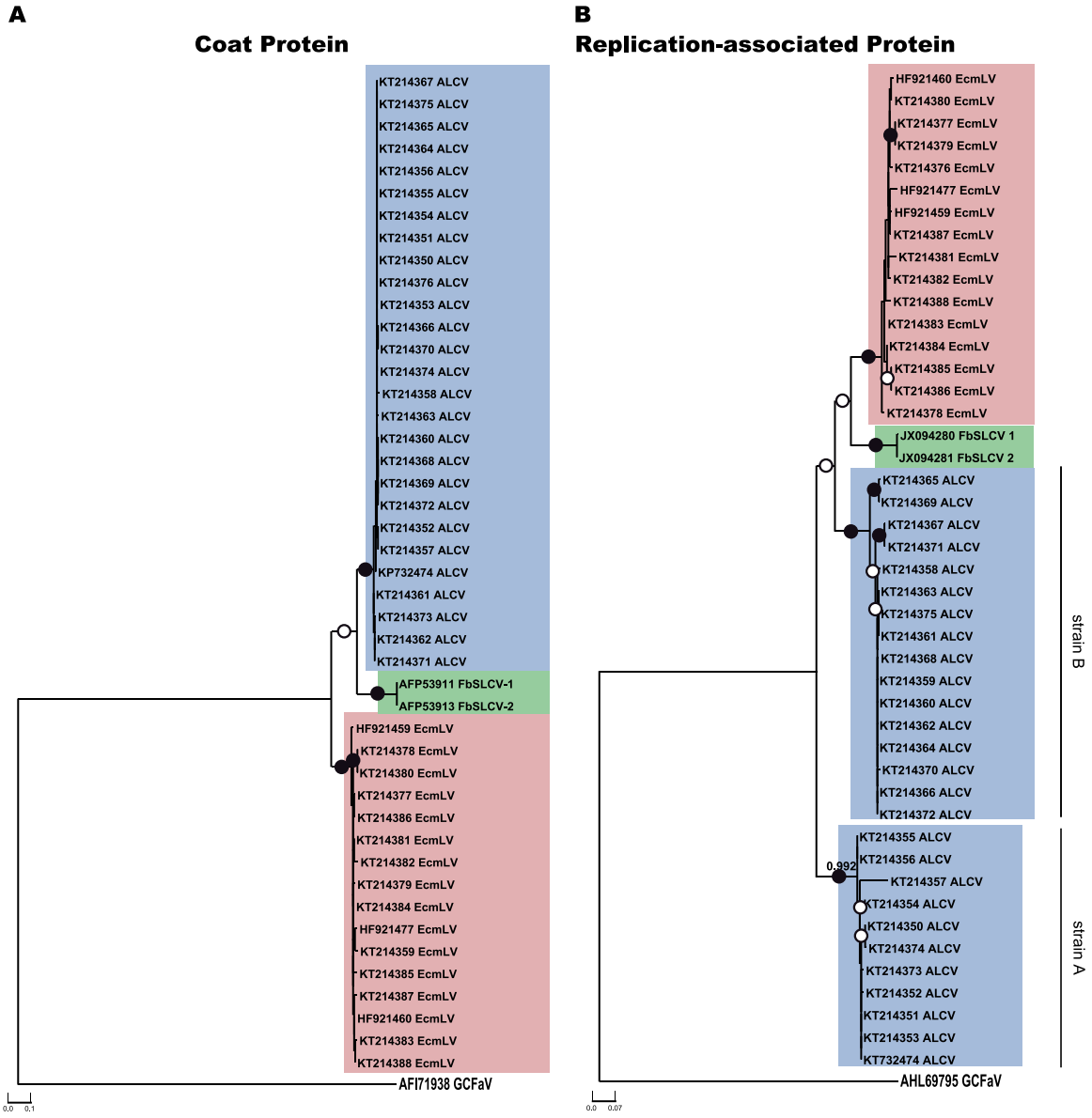

Fig. 3. (A) and (B) Maximum-likelihood phylogenetic trees of, respectively predicted CP and Rep amino acid sequences of 45 isolates of the three capulavirus species. Branches associated with a filled dot have bootstrap supports above 85% whereas those with an unfilled dot have bootstrap supports above 50%. The maximum-likelihood phylogenetic trees (A) and (B) are rooted with the Grapevine cabernet franc-associated virus (GCFaV, AHL69795) Rep and (GCFaV, AFI71938) CP, respectively.

Table 1
Recombination events detected in capulaviruses.

Event	Recombinant(s)	Major parent	Minor parent	Methods[a]	Breakpoints positions[b]
1	ALCV_VAU14_LUZ142	ALCV_ASS34_Assas1	Unknown	RGBMC	1433-2708
2	ALCV_PB14_LUZ178	ALCV_ASS34_Assas1	ALCV_PB14_LUZ171	RGBMCST	866-1046
3	ALCV_PB14_GS6	Unknown	ALCV_ASS14_Assas1	RGBMCST	1357–2112
4	ALCV_PB14_LUZ165	Unknown	ALCV_ASS14_Assas1	RGBMCST	2757–908
5	ALCV_PB14_LUZ171	Unknown	ALCV_PB14_LUZ184	RGBMCST	1033 (nad)[c] – 1716
6	ALCV_PB14_LUZ184	ALCV_GAG13_LUZ193	EcmLV_CM251	RGBM	884–969
7	ALCV_ASS14_Assas2	ALCV_GAG13_LUZ193	ALCV_TDV12_48-2 A	RGMCT	3011 (nad) – 346 (nad)
8	ALCV_PB14_GS4	ALCV_GAG13_LUZ193	Unknown	RMST	2375 (nad) – 642 (nad)
9	ALCV_PB14_LUZ178	EcmLV_MP4C	Unknown	RBMCS	531–1009 (nad)
10	ALCV_PB14_LUZ171	ALCV_PB13_LUZ165	Unknown	RGMC	2382–2726 (nad)
11	EcmLV_CM243	Unknown	FbSLCV-1	GBC	2366–2652 (nad)
12	ALCV_VAU14_LUZ136	ALCV_PB13_LUZ188	Unknown	RGBMCST	451 (nad) – 1599 (nad)
13	ALCV_ASS14_Assas1	ALCV_PB13_LUZ188	Unknown	GBMS	1196 (nad) – 1599 (nad)
14	ALCV_PB14_LUZ188	ALCV_TDV13_48-2 A	ALCV_ALB14_LUZ147	RGM	451 (nad) – 1599 (nad)

[a] RDP (R), GENECONV (G), BOOTSCAN (B), MAXIMUM CHI SQUARE (M), CHIMAERA (C), SISCAN (S) and 3SEQ (T) recombination detection methods.
[b] Begin and end breakpoints positions in the recombinant sequence.
[c] nad: not accurately determined.

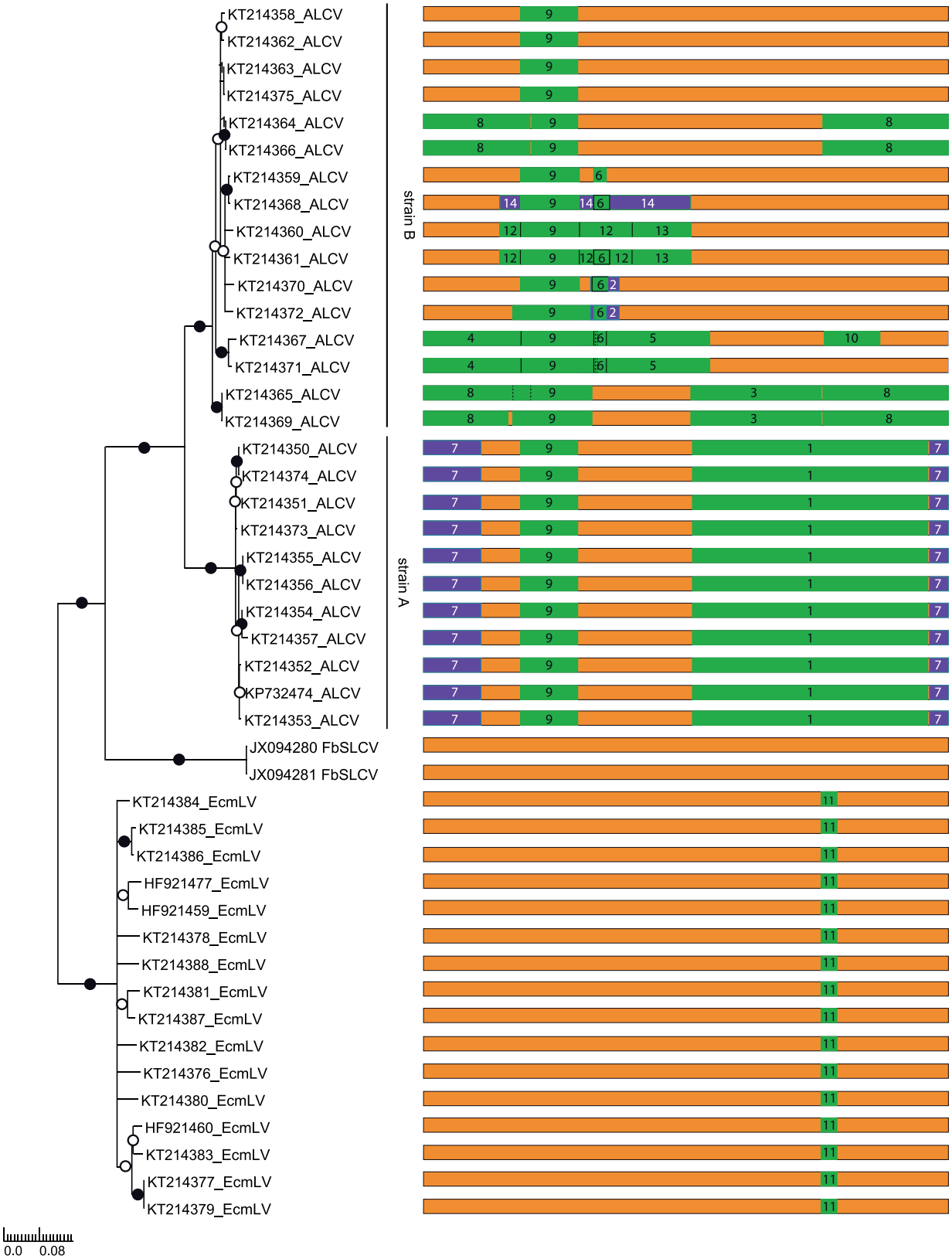

Fig. 4. Maximum-likelihood phylogenetic tree depicting the relatedness of the non-recombinant genomic regions of 45 isolates of the three capulavirus species. The fourteen unique recombination events detected within these sequences are presented to the right of the tree. Green and purple colors indicate genome regions that have likely been acquired by intra- and inter-species sequence transfers. Recombination events are numbered according to Table 1. Branches associated with a filled dot have bootstrap supports above 90% whereas those with an unfilled dot have bootstrap supports above 70%. All branches with less than 50% bootstrap support have been collapsed.

(i.e. EcmLV, FbSLCV or ALCV; Table 1) and another involving either an undescribed capulavirus species, or even perhaps a virus from an undescribed geminivirus genus.

Events 9 and 11 are respectively shared by all ALCV and EcmLV isolates (Supplementary Table 4), which indicates that these two events predate the most recent common ancestors of the analyzed ALCVs and EcmLVs and are, therefore, more ancient than the rest of the identified events. In addition, inter-species recombination event 1 and inter-ALCV strain recombination event 7 may account for the clear divergence of the ALCV-A and -B strains from one another.

3.6. Comparisons between the spatial distribution and genetic diversity of ALCV and EcmLV

Mantel tests were used to determine whether there was any correlation between the genetic and geographical distances of the sampled isolates. For ALCV, no significant correlation was observed between geographic distance and genetic distance for the fragment encompassing the V3 ORF and part of the *cp* ORF (Mantel R correlation scores of −0.034 with an associated *p*-value=0.230; Supplementary Fig. 5). This result suggests free movement of ALCV across the French Mediterranean region.

It is interesting that symptoms in alfalfa plants that resemble in some respect those described here (including plant stunting and leaf curling, crumpling and shriveling) have been described throughout the Mediterranean basin (France, Bulgaria, Romania, Spain and Saudi Arabia) from the late 1950s to the 1980s (Alliot et al., 1972; Blattný, 1959; Cook and Wilton, 1984; Leclant et al., 1973; Rodriguez Sardiña and Novales Lafarga, 1973) and, more recently, in Argentina (Bejerman et al., 2011). However, enations, which were associated with alfalfa leaf curling in these reports, have not been observed in ALCV infected plants either in the field or the laboratory. Attempts to identify viral particles by electron microscopy in some of the earliest reports revealed that the tissues within enations contained bullet-shaped, rhabdovirus-like particles with the viral species producing these particles being tentatively named Lucerne enation virus (LEV) (Alliot et al., 1972; Rodriguez Sardiña and Novales Lafarga, 1973). Additional experiments indicated that both grafting and *A. craccivora* (but not *Acyrthosiphon pisum*) transmission could successfully spread LEV symptoms between alfalfa plants. On the other hand, mechanical and leafhopper (*Calligypona pellucida* F.) transmission failed (Alliot et al., 1972; Blattný, 1959; Leclant et al., 1973; Rodriguez Sardiña and Novales Lafarga, 1973). It was therefore concluded that LEV was a circulatively transmitted virus within the family *Rhabdoviridae* (Alliot et al., 1972). However, because differences in the types of symptoms observed depended on the mode of transmission, Rodriguez Sardiña and Novales Lafarga (1973) hypothesized that another virus might frequently be present in co-infection with LEV (Rodriguez Sardiña and Novales Lafarga, 1973). To test if, apart from *Alfalfa mosaic virus*, which is ubiquitous in alfalfa, another virus could be present in co-infections, we examined four rhabdovirus-infected alfalfa plants collected from Spain and discovered that all of them were indeed co-infected with ALCV. Besides indicating that the geographical range of ALCV extends beyond the borders of France, this result is consistent with the hypothesis of Sardiña and Lafarga: i.e. that the disease attributed to LEV in the 1970s could potentially be caused by a complex of two or more viruses, one of which we now know is likely to be ALCV. This hypothesis will need to be further tested by examining further symptomatic alfalfa plants from around the Mediterranean basin.

For EcmLV, geographic distances between sampling locations were significantly correlated with genetic distances between the *rep* gene fragments of the viral isolates (Mantel *r* correlation score=0.142, *p*-value=0.002; Supplementary Fig. 5). This indicates that a degree of differentiation exists between EcmLV populations at a sub-regional scale within the Western Cape, which in turn suggests that there are likely restrictions on the free movement of EcmLV across the region.

The observed differences between the spatial distribution and genetic diversity of ALCV and EcmLV might reflect general differences between viruses that infect cultivated and non-cultivated hosts. Relative to viruses such as ALCV that infect cultivated hosts, the population genetic structures of viruses such as EcmLV that preferentially infect uncultivated hosts are likely to be impacted by a more complex combination of biotic parameters. Variable distributions and population densities of suitable host plants within natural environments can influence the probability that viruliferous insect vectors will successfully transmit viruses to an appropriate host (Keesing et al., 2006). Also, the variable life-spans of host plants, the possibility of long-term vertical transmission chains when hosts are vegetatively propagated, and the potential for sporadic vector transmission will all contribute to the selection processes that ultimately shape the population genetic structure of viruses that are adapted to infecting uncultivated species such as *E. caput-medusae*. It is entirely plausible, however, that, as with ALCV, EcmLV is also transmitted by *A. craccivora* (which is polyphagous, very widely distributed and has even been repeatedly observed on *E. caput-medusae*; Supplementary Fig. 6), further studies are needed to test this hypothesis.

4. GenBank accession numbers

Full genomes of: Alfalfa leaf curl virus (KT214350-KT214375); Euphorbia caput-medusae latent virus (KT214376-KT214388). V3 ORF and part of the *cp* ORF of Alfalfa leaf curl virus (KT214391-KT214427). C3 ORF and part of the C1 ORF of Euphorbia caput-medusae latent virus (KT964062-KT964084).

Acknowledgments

We wish to express our sincere thanks and appreciation to Mr Paul Loubser and colleagues from Buffelsfontein Game & Nature Reserve and to Mrs. Hestelle Melville and Miss Laurenda Van Breda from the University of the Western Cape Nature Reserve Unit. We also thank Michel Peterschmitt for helpful discussions and Stéphane Blanc for effective manuscript review. DPM, AV and GWH are supported by the National Research Foundation of South Africa (Grant N° TTK1207122745). PH is supported by the Poliomyelitis Research Foundation (Grant N° 15/102). This work was supported by Direction Générale de l'Armement (Grant N° 201160060) (Ministère de la Défense, France), The Métaprogramme INRA «Meta-omics of microbial ecosystems» (Grant N° 24000466) and CIRAD.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2016.03.016.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Agindotan, B.O., Domier, L.L., Bradley, C.A., 2015. Detection and characterization of the first North American mastrevirus in switchgrass. Arch. Virol. 160, 1313–1317.

Alliot, B., Signoret, P.A., Giannott., J., 1972. Presentation of bacilliform virus-particles associated with enation disease of alfalfa (Medicago-Sativa L). Cr Acad. Sci. D Nat. 274, 1974–1976.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Arguello-Astorga, G.R., Guevara-Gonzalez, R.G., Herrera-Estrella, L.R., Rivera-Bustamante, R.F., 1994. Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. Virology 203, 90–100.
Bejerman, N., Nome, C., Giolitti, F., Kitajima, E., de Breuil, S., Pérez Fernández, J., Basigalup, D., Cornacchione, M., Lenardon, S., 2011. First report of a rhabdovirus infecting alfalfa in Argentina. Plant Dis. 95 771-771.
Bernardo, P., Golden, M., Akram, M., Naimuddin, Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A.G., Peterschmitt, M., Martin, D.P., Roumagnac, P., 2013. Identification and characterisation of a highly divergent geminivirus: evolutionary and taxonomic implications. Virus Res. 177, 35–45.
Blattný, C., 1959. Virus papillosity of the leaves of lucerne. Folia Microbiol. 4, 212–215.
Briddon, R.W., Heydarnejad, J., Khosrowfar, F., Massumi, H., Martin, D.P., Varsani, A., 2010. Turnip curly top virus, a highly divergent geminivirus infecting turnip in Iran. Virus Res. 152, 169–175.
Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J. C., Fiallo-Olive, E., Briddon, R.W., Hernandez-Zepeda, C., Idris, A., Malathi, V.G., Martin, D.P., Rivera-Bustamante, R., Ueda, S., Varsani, A., 2015. Revision of Begomovirus taxonomy based on pairwise sequence comparisons. Arch. Virol. 160, 1593–1619.
Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J.H., Fernandez, E., Martin, D.P., Varsani, A., Roumagnac, P., 2014. Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. Plos. One 9, e102945.
CIE, 1983. Distribution Maps of Plant Pests, in: Proceedings of the CAB International, N.W., Wallingford, Oxfordshire, OX10 8DE, UK. (Ed.). CAB International Wallingford UK.
Cook, A.A., Wilton, A.C., 1984. Alfalfa enation virus in the Kingdom of Saudi Arabia. FAO Plant Prot. Bull. 32, 139–140.
Dry, I.B., Rigden, J.E., Krake, L.R., Mullineaux, P.M., Rezaian, M.A., 1993. Nucleotide sequence and genome organization of tomato leaf curl geminivirus. J. Gen. Virol. 74, 147–151.
Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf. 5, 113.
Filloux, D., Dallot, S., Delaunay, A., Galzi, S., Jacquot, E., Roumagnac, P., 2015a. Metagenomics approaches based on virion-associated nucleic acids (VANA): an innovative tool for assessing without a priori viral viversity of plants. Methods Mol. Biol. 1302, 249–257.
Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., Uzest, M., Rodier-Goud, M., D'Hont, A., Vernerey, M.S., Wilkin, P., Peterschmitt, M., Winter, S., Murrell, B., Martin, D.P., Roumagnac, P., 2015b. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus Evol. 1, 1–17.
Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.
Golden, M., Martin, D., 2013. DOOSS: a tool for visual analysis of data overlaid on secondary structures. Bioinformatics 29, 271–272.
Guindon, S., Delsuc, F., Dufayard, J.F., Gascuel, O., 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537, 113–137.
Gutierrez, C., Ramirez-Parra, E., Mar Castellano, M., Sanz-Burgos, A.P., Luque, A., Missich, R., 2004. Geminivirus DNA replication and cell cycle interactions. Veter. Microbiol. 98, 111–119.
Haible, D., Kober, S., Jeske, H., 2006. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J. Virol. Methods 135, 9–16.
Inoue-Nagata, A.K., Albuquerque, L.C., Rocha, W.B., Nagata, T., 2004. A simple method for cloning the complete begomovirus genome using the bacteriophage phi 29 DNA polymerase. J. Virol. Methods 116, 209–211.
Keesing, F., Holt, R.D., Ostfeld, R.S., 2006. Effects of species diversity on disease risk. Ecol. Lett. 9, 485–498.
Kraberger, S., Farkas, K., Bernardo, P., Booker, C., Arguello-Astorga, G.R., Mesleard, F., Martin, D.P., Roumagnac, P., Varsani, A., 2015. Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Arch. Virol. 160, 1303–1311.
Krenz, B., Thompson, J.R., Fuchs, M., Perry, K.L., 2012. Complete genome sequence of a new circular DNA virus from grapevine. J. Virol. 86, 7715.
Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1–7.
Lazarowitz, S.G., 1992. Geminiviruses - genome structure and gene-function. Crit. Rev. Plant Sci. 11, 327–349.
Lazarowitz, S.G., Pinder, A.J., Damsteegt, V.D., Rogers, S.G., 1989. Maize streak virus genes essential for systemic spread and symptom development. EMBO J. 8, 1023–1032.
Leclant, F., Alliot, B., Signoret, P.A., 1973. Transmission et épidémiologie de la maladie à énations de la luzerne (LEV). Premiers résultats. Ann. Phytopathol. 5, 441–445.
Li, S.C., Shiau, C.K., Lin, W.C., 2008. Vir-Mir db: prediction of viral microRNA candidate hairpins. Nucleic Acids Res. 36, D184–D189.
Liang, P., Navarro, B., Zhang, Z., Wang, H., Lu, M., Xiao, H., Wu, Q., Zhou, X., Di Serio, F., Li, S., 2015. Identification and characterization of a novel geminivirus with monopartite genome infecting apple trees. J. Gen. Virol. 96, 2411–2420.
Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G.P., Saponari, M., 2012. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease a new member in the family Geminiviridae. Virology 432, 162–172.
Ma, Y., Navarro, B., Zhang, Z., Lu, M., Zhou, X., Chi, S., Di Serio, F., Li, S., 2015. Identification and molecular characterization of a novel monopartite geminivirus associated with mulberry mosaic dwarf disease. J. Gen. Virol. 96, 2421–2434.
Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.
Markham, N.R., Zuker, M., 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. 453, 3–31.
Martin, D.P., Murell, B., Golden, M., Khoosal, A., Muhire, B., 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 1, vev003.
Muhire, B., Martin, D.P., Brown, J.K., Navas-Castillo, J., Moriones, E., Zerbini, F.M., Rivera-Bustamante, R., Malathi, V.G., Briddon, R.W., Varsani, A., 2013. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Arch. Virol. 158, 1411–1424.
Muhire, B.M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J.M., Gray, A., Poon, A.Y., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.L., Harkins, G.W., Varsani, A., Shepherd, D.N., Martin, D.P., 2014a. Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. J. Virol. 88, 1972–1989.
Muhire, B.M., Varsani, A., Martin, D.P., 2014b. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. Plos One 9, e108277.
Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Pond, S.L.K., Scheffler, K., 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. 30, 1196–1205.
Ng, T.F., Chen, L.F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P.D., Varsani, A., Kondov, N.O., Wong, W., Deng, X., Andrews, T.D., Moorman, B.J., Meulendyk, T., MacKay, G., Gilbertson, R.L., Delwart, E., 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc. Natl. Acad. Sci. USA 111, 16842–16847.
Ng, T.F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L., Breitbart, M., 2009. Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. J. Virol. 83, 2500–2509.
Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. Plos One 6, e19050.
Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., Ohta, H., 2007. ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res. 35, D863–D869.
Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.
Poon, A.F., Lewis, F.I., Frost, S.D., Kosakovsky Pond, S.L., 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 24, 1949–1950.
Rodriguez Sardiña, J., Novales Lafarga, J., 1973. Una virosis de la alfalfa con producción de "enations". An. INIA/Ser. Prot. veg. 3, 132–146.
Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant virus metagenomics: advances in virus discovery. Phytopathology 105, 716–727.
Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 1–9.
Rosario, K., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. Arch. Virol. 157, 1851–1871.
Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E.J., Collings, D.A., Walters, M., Martin, D.P., Breitbart, M., Varsani, A., 2011. Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). J. Gen. Virol. 92, 1302–1308.
Rosario, K., Seah, Y.M., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J.E., Duffy, S., Breitbart, M., 2015. Vector-Enabled Metagenomic (VEM) surveys using whiteflies (aleyrodidae) reveal novel begomovirus species in the new and oldworlds. Viruses 7, 5553–5570.
Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez, E., Julian, C., Abt, I., Filloux, D., Mesleard, F., Varsani, A., Blanc, S., Martin, D.P., Peterschmitt, M., 2015. Alfalfa leaf curl virus: an aphid-transmitted geminivirus. J. Virol. 89, 9683–9688.
Schubert, J., Habekuss, A., Kazmaier, K., Jeske, H., 2007. Surveying cereal-infecting geminiviruses in Germany–diagnostics and direct sequencing using rolling circle amplification. Virus Res. 127, 61–70.
Schubert, S., Grunweller, A., Erdmann, V.A., Kurreck, J., 2005. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. J. Mol. Biol. 348, 883–893.
Semegni, J.Y., Wamalwa, M., Gaujoux, R., Harkins, G.W., Gray, A., Martin, D.P., 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27, 2443–2445.
Shepherd, D.N., Martin, D.P., Lefeuvre, P., Monjane, A.L., Owor, B.E., Rybicki, E.P., Varsani, A., 2008. A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. J. Virol. Methods 149, 97–102.
Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.
Theiler, J., Bloch, J., 1996. Nested test for point sources. In: Babu, G.J., Feigelson, E.D. (Eds.), Statistical Challenges in Modern Astronomy II. Springer-Verlag, New York, pp. 407–408.

Thompson, M.D., Jacks, C.M., Lenvik, T.R., Gantt, J.S., 1992. Characterization of rps17, rp19 and rpl15: three nucleus-encoded plastid ribosomal protein genes. Plant Mol. Biol. 18, 931–944.
Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Murilo Zerbini, F., Brown, J.K., 2014a. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch. Virol. 159, 1873–1882.
Varsani, A., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Brown, J. K., Murilo Zerbini, F., Martin, D.P., 2014b. Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Arch. Virol. 159, 2193–2203.
Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P., Martin, D.P., 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. Virol. J. 6, 36.
Yazdi, H.R.B., Heydarnejad, J., Massumi, H., 2008. Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel non-anucleotide. Virus Genes 36, 539–545.